AD_____

Award Number: DAMD17-00-1-0448

TITLE: Cox Model for Interval Censored Data in Breast Cancer Follow-up Studies

PRINCIPAL INVESTIGATOR: George Wong, Ph.D.

CONTRACTING ORGANIZATION: Strang Cancer Prevention Center
New York, New York 10021-4601

REPORT DATE: July 2001

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

20011127 095

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 2001	Annual (1 Jul 00 - 30 Jun 01)

4. TITLE AND SUBTITLE
Cox Model for Interval Censored Data in Breast Cancer Follow-up Studies

5. FUNDING NUMBERS
DAMD17-00-1-0448

6. AUTHOR(S)
George Wong, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Strang Cancer Prevention Center
New York, New York 10021-4601

E-Mail: gwong@strang.org

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT: The overall objective of this research proposal is semi-parametric inference of the Cox regression model for a survival function $\Pr(X > x|Z = z) = S(x|z) = [S_0(x)]^{e^{z\beta}}$, where X is subject to interval censoring, Z represents the covariates, S_0 is a baseline survival function, and β represents the regression coefficients. One objective of our research is to develop asymptotic inference of the generalized maximum likelihood estimator (GMLE) of the regression coefficients β and $S(\cdot|z)$. A critical limitation with the GMLE approach under interval censoring is that it is computationally feasible only for a small data set. Thus the focus of another aspect of our research is the investigation of a simple alternative to the GMLE obtained by a two-step estimation procedure involving data grouping. In the first year of our research, we have completed a computer program for iterative calculating the two-step estimator (TSE) of β and $S(s|z)$. We have demonstrated by Monte Carlo simulations that the TSE is robust against degree of partitioning in data grouping; moreover, the TSE is consistent. The results will be useful to breast cancer researchers pursuing chemoprevention intervention trials involving surrogate endpoint biomarkers, and genetic epidemiologists conducting studies on familial aggregation of breast cancer and related cancers.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Breast cancer, interval-censored data, Cox regression model, generalized maximum likelihood, two-step estimation, consistency, asymptotic normality and efficiency	11
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

___ Where copyrighted material is quoted, permission has been obtained to use such material.

___ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

___ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

N/A For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

A. TABLE OF CONTENTS

Front Cover	1
Report Documentation Page	2
Forward	3
A. Table of contents	4
B. Introduction	5 – 6
C. Body	6 – 9
D. Key research accomplishments in the first year	10
E. Reportable outcomes	10
F. Conclusions	10
G. References	10 – 11

B. INTRODUCTION

Interval-censored (IC) data are encountered in three areas of breast cancer research. The most common application is in clinical relapse follow-up studies in which the study endpoint is disease-free survival. When a patient relapses, it is usually known that the relapse takes place between two follow-up visits, and the exact time to relapse is unknown. In statistics, we say relapse time is interval censored. Interval censoring is also encountered in breast cancer registry studies in which information on family history of cancer is updated periodically. The Strang Breast Surveillance Program for women at increased risk for breast cancer, for instance, has enlisted over 800 women with complete pedigree information which is verified and updated continuously. Family history data such as age at diagnosis of a specific cancer, or a benign but risk-conferring condition, are obtained from each registrant at each update. Time to a cancer event, and definitely time to first detection of a benign condition, are at best known to fall in the time interval between the last update and age at diagnosis. A third but increasingly important area of application of interval censoring is in breast cancer chemoprevention experiments or prevention trials, which involve the observation of one or more surrogate endpoint biomarkers (SEB) over time. The scientific question of interest here is the estimation of time for the SEB to reach a target value, and time from cessation of intake of a chemopreventive agent to the loss of its protective effect. Unfortunately, the exact values of both these time variables are known only to lie in between two successive assay inspection times. In a breast cancer follow-up study, we will often encounter covariates (for instance, tumor size and nodal status in a relapse study, and baseline SEB value in a chemoprevention trial).

Let X denote a time-to-event variable with distribution $F(x) = Pr(X \leq x)$, or equivalently, survival function $S(x) = 1 - F(x)$. In interval censoring, X is not observed and is known only to lie in an observable interval (L, R). In our previous DOD funded grant, we have made fundamental contributions to both the theory of the generalized maximum likelihood (GML) estimation of S, and the computation in connection with the inference of GML estimator (GMLE) $\hat{S}$ of S. These contributions are restricted to the case of univariate interval-censored data without covariates.

The Cox proportional hazards model [1] specifies that covariates have a proportional effect on the hazard function of X. This model provides powerful means for fitting failure time observations to a distribution free model and for estimating the risk for failure associated with a vector of covariates. It is extensively used for right-censored data. Finkelstein [2] applied the Cox model to analysis of interval-censored data. However, she did not establish asymptotic properties of the GMLE of the parameters in the model and the approach is limited to small sample sizes due to the computational difficulty .

Our interest in IC data with covariates is driven by needs arising from two related areas of breast cancer research at Strang. First, our investigators in the Strang Cancer Genetics Program want to study various patterns of familial aggregation of breast, ovarian and other forms of cancer using family history data from the Strang Breast Surveillance Program. Studies of familial early onset of breast cancer, breast-ovarian and breast-prostate associations will lead to IC data with covariates; therefore, a proper statistical procedure together with a feasible software to deal with such data are very much needed. Second, we conducted a one-year chemoprevention trial of indole-3-carbinol (I3C) for breast cancer

prevention. In this prevention trial we monitored the levels of two SEB's, a urinary estrogen metabolite ratio and a blood counterpart, both of which are subject to interval censoring. An earlier dose-ranging study of I3C conducted by Wong *et al* [2] has been published.

The overall aim of this research proposal is to develop statistical inference for interval-censored data with covariates that are encountered in breast cancer chemoprevention trials employing surrogate endpoint biomarkers, and in breast cancer registry follow-up studies of familial aggregation of breast and other forms of cancer. Asymptotic generalized maximum likelihood theory under the Cox regression model will be investigated and computer software package for maximum likelihood inference will be implemented.

C. BODY

C.1. Model Formulation and Likelihood Equations.

Let $Y_{K,1} < Y_{K,2} < \cdots < Y_{K,K}$ denote the follow-up times for a patient who has made K follow-up visits, in a longitudinal follow-up study. Since the number of visits for each patient may vary, K is a random positive integer. For convenience, define $Y_{K,0} = 0$ and $Y_{K,K+1} = \infty$. The time-to-event variable of interest, X, is not directly observed; instead, it is known to lie in between two successive censoring time points $(Y_{K,j}, Y_{K,j+1})$, where j =0, ..., K. Note that X is left censored if $j = 0$, strictly interval censored if $0 < j < K$, and right censored if $X > Y_{K,K}$. The observable interval-censored data corresponding to X is given by

$$(L, R) = (Y_{K,i}, Y_{K,i+1}) \text{ if } Y_{K,i} < X \leq Y_{K,i+1},\ i = 0, 1, ..., K. \tag{2.1}$$

In addition to (L, R), we also observe a $p \times 1$ covariate vector Z. We assume that K and the $Y_{k,j}$'s are independent of (X, Z).

The Cox regression model for the survival function at $X = x$ given $Z = z$ is represented by

$$S(x|z) = [S_0(x)]^{e^{z\beta}},$$

where $z\beta$ is the dot product of Z and β, $S_0(x)$ is a baseline survival function and β is a p-dimensional regression coefficient vector.

Let $I_i = (L_i, R_i, z_i)$, i = 1, ..., n, be a random sample of size n interval-censored observations with covariates. In terms of the original observed intervals, the likelihood function of S and b is given by

$$\mathrm{L} = \prod_{i=1}^{n}((S(L_i))^{e^{bz_i}} - (S(R_i))^{e^{bz_i}}), \tag{2.2}$$

where S is a survival function, and b is a $p \times 1$ dimensional vector. The GMLE of (S_0, β) is a value (S, b) that maximizes (2.2) over all survival functions S and all $b \in \mathcal{R}^p$.

Since S_o places all probability mass on the innermost intervals of the I_i's (see Peto (1973) or Turnbull (1976)), it is often computationally simpler to express L in terms of innermost intervals.

We say that an interval A is an innermost interval of the I_i's if A is a nonempty finite intersection of one or more of the I_i's such that either $I_i \cap A = \emptyset$ or $I_i \cap A = A$ for each

i. Suppose there are a total of m distinct innermost intervals $A_i = (\xi_i, \eta_i]$, where $\eta_i \leq \xi_{i+1}$ and $m \leq n$. Then the likelihood function (2.2) is equivalently given by

$$\text{Ł} = \prod_{i=1}^{n} [(\sum_{k>l_i} s_k)^{e^{z_i b}} - (\sum_{k>r_i} s_k)^{e^{z_i b}}], \tag{2.3}$$

where $l_i = \sup\{j : \eta_j \leq L_i\}$, $r_i = \sup\{j : \eta_j \leq R_i\}$ and $s = (s_1, ..., s_m)$ denote the vector of the probability weights. The log likelihood of (s, b) is

$$\mathcal{L}(s, b) = \sum_{i=1}^{n} \ln[(\sum_{k>l_i} s_k)^{e^{z_i b}} - (\sum_{k>r_i} s_k)^{e^{z_i b}}]. \tag{2.4}$$

Note that $(\sum_{k>r_i} s_k)^{e^{z_i b}} = 1$ if $r_i = 0$ and $(\sum_{k>l_i} s_k)^{e^{z_i b}} = 0$ if $l_i = m$.

C.2. Generalized maximum likelihood estimation.

A GMLE of (s, β) is a value of (s, b) that maximizes the likelihood function (2.4). We could follow the Newton-Raphson (NR) algorithm taken by Finkelstein [2]. However, this would involve the inverse of a matrix of order $(m + p - 1) \times (m + p - 1)$. Since m can be potentially large when n is large, the NP algorithm is not feasible for a large data set. In our simulation studies with $n = 200$, m ranges from 17 to 22.

We advocate a computationally simple approach by first grouping the original data (L_i, R_i) and then applying a two-step iterative scheme to obtain the two-step estimators (TSE) of s_o and β based on the innermost interval corresponding to the grouped intervals.

In the first year of our research, we have successfully implemented the computer software to calculate the TSE's of s_o and β. The algorithm is summarized as follows.

1. Partition the entire data range into q time points, $q < n$. Let (L_i^*, R_i^*) denote the grouped observable intervals, $i = 1, ..., n$. Let $\underline{s} = (s_1, ..., s_{m_q})$ denote the vector of probability masses distributed over the $m_q < m$ innermost intervals corresponding to the (L_i^*, R_i^*)'s
2. Maximize the likelihood of $\underline{s}$ and $\underline{b}$ based on the grouped data using a two-step maximization algorithm. At each iteration of the algorithm, there is an $\underline{s}-$step in which the likelihood is increased by changing a transformed parameter of $\underline{s}$, while $\underline{b}$ is fixed at the value from the previous iteration. This is followed by a $\underline{b}-$step in which the likelihood is maximized with respect to $\underline{b}$ with $\underline{s}$ fixed at the value updated at the current $\underline{s}-$step.

For ease of presentation, we outline the algorithm in the case $m_q = 3$ so that $\underline{s} = (s_1, s_2, s_3)$.

$\underline{s}-$step:

a. Transform $\underline{s}$ to $\underline{s}(u)$, where $\underline{s}(u) = (s_1(u), s_2(u), s_3(u))$,

$$s_1(u) = \frac{s_1 + u}{1 + u}, \; s_2(u) = \frac{s_2}{1 + u}, \; s_3(u) = \frac{s_3}{1 + u},$$

and u is such that $u + s_1 > 0$.

b. Use NP algorithm to maximize $\mathcal{L}(\underline{s}(u), \underline{b})$ with respect to u. Denote the maximizer by u_1. Let $\underline{s}^* = \underline{s}(u_1)$.

c. Transform $\underline{s}^*$ to $\underline{s}^*(u)$, where $\underline{s}^*(u) = (s_1^*(u), s_2^*(u), s_3^*(u))$,

$$s_1^*(u) = \frac{s_1^*}{1+u},\ s_2^*(u) = \frac{s_2^* + u}{1+u},\ \text{and}\ s_3^*(u) = \frac{s_3^*}{1+u}.$$

d. Use NP algorithm to maximize $\mathcal{L}(\underline{s}^*(u), \underline{b})$ with respect to u. Denote the maximizer by u_2. Let $\underline{s}^{**} = \underline{s}^*(u_2)$.

e. Transform $\underline{s}^{**}$ to $\underline{s}^{**}(u)$, where $\underline{s}^{**}(u) = (s_1^{**}(u), s_2^{**}(u), s_3^{**}(u))$,

$$s_1^{**}(u) = \frac{s_1^{**}}{1+u},\ s_2^{**}(u) = \frac{s_2^{**}}{1+u},\ \text{and}\ s_3^{**}(u) = \frac{s_3^{**} + u}{1+u}.$$

f. Use NP algorithm to maximize $\mathcal{L}(\underline{s}^{**}(u), \underline{b})$ with respect to u. Denote the maximizer by u_3. Let $\underline{s}^{***} = \underline{s}^{**}(u_3)$.

$\underline{b}$–step:

Use NP algorithm to maximize $\mathcal{L}(\underline{s}^{***}, \underline{b})$ with respect to $\underline{b}$.

Repeat $\underline{s}$–step and $\underline{b}$–step until convergence.

C.3. Sensitivity study of TSE.

We have carried out Monte Carlo simulations to investigate the sensitivity of the TSE of $\underline{\beta}$ to the degree of partitioning used in the data grouping. Our simulation studies are designed as follows:

1. X is exponential with pdf $f(x) = \frac{1}{\sigma} e^{-[\frac{x-\mu}{\sigma}+1]} \mathbf{1}[(x > \mu - \sigma)]$, where $\mathbf{1}[\cdot]$ denotes the indicator function.
2. There are 3 mutually independent covariates Z_1, Z_2 and Z_3, each of which is a discrete random variable with pdf $f(i) = \frac{i}{\sum_{j=1}^{6} j}$, $i = 1, ..., 6$.
3. (L, R) is generated according to the following scheme:

$$(L, R) = \begin{cases} (0, U) & \text{if } X \le U, \\ (20, \infty) & \text{if } X > 20, \\ (U + kV, U + (k+1)V) & \text{if } X \le 20,\ U + kV < X \le U + (k+1)V \text{ and } k \ge 1. \end{cases}$$

where $U \sim U(0, 2)$ and $V \sim U(0, 2.3)$.

For each of Monte Carlo simulation, a total of 1000 replications are performed. Grouping width of sizes 3, 5 and 8 are considered in the partitioning of the interval $[0, 20]$.

Tables 1 and 2 summarize the simulation results for sample sizes $n = 30$ and $n = 200$, respectively. For original data (no grouping), GMLE values are given. In the 3 cases of data grouping, TSE values are listed. The figures given in parentheses are standardized differences

defined as $\frac{|\text{ sample mean of estimator - true value }|}{\text{sample standard error of estimator}}$. Essentially, the TSE is quite insensitive to changes in the degree of partitioning, and sample size appears to be not a relevant issue either. The conclusion here applies to the parameter estimate of $\underline{\beta}$ only. However, in assessing closeness of asymptotic inference of the TSE to that of the GMLE, we will have to pay attention to the asymptotic covariance matrix of the TSE of $\underline{\beta}$. Because the covariance matrix will be a function of the probability weights s_1, ..., s_{m_g}, it is clear that the degree of partitioning can affect the asymptotic approximation of the TSE to the GMLE in a more significant way. We will relegate this aspect of research to the second year of our DOD grant.

data	cpu time	$\beta_1 = -0.1$	$\beta_2 = 0.2$	$\beta_3 = -0.1$
original	10 min.	-0.092 (0.083)	0.209 (0.085)	-0.092 (0.081)
grouped width=3	4 min.	-0.093 (0.068)	0.217 (0.136)	-0.094 (0.055)
grouped width=5	4 min.	-0.094 (0.050)	0.239 (0.219)	-0.094 (0.048)
grouped width=8	*	*	*	*

* Calculation not possible due to sparseness of data

Table 1. Monte Carlo simulations for TSE of $\underline{\beta}$ for $n = 30$

data	cpu time	$\beta_1 = -0.1$	$\beta_2 = 0.2$	$\beta_3 = -0.1$
original	58.3 min.	-0.087 (0.419)	0.191 (0.281)	-0.087 (0.406)
grouped width=3	6.5 min.	-0.089 (0.333)	0.198 (0.057)	-0.089 (0.333)
grouped width=5	5.6 min.	-0.090 (0.294)	0.203 (0.077)	-0.090 (0.286)
grouped width=8	4.7 min.	-0.092 (0.205)	0.212 (0.250)	-0.093 (0.171)

Table 2. Monte Carlo simulations for TSE of $\underline{\beta}$ for $n = 200$

Incidentally, the close approximation of the GMLE values of $\underline{\beta}$ to the true $\underline{\beta}$ (first row of Table 1 or Table 2) indicates that the GMLE of $\underline{\beta}$ is consistent. Similarly, rows 2-4 of Table 1 or Table 2 suggest that the TSE of $\underline{\beta}$ is consistent.

D. KEY RESEARCH ACCOMPLISHMENTS IN THE FIRST YEAR

- We have completed Task 1.
  We have successfully implemented a computer progam to calculate the TSE of the Cox regression coefficients $\underline{\beta}$.
- We have completed Task 2.
  We have demonstrated by Monte Carlo simulations that the TSE of $\underline{\beta}$ is not much affected by the degree of partitioning used in data grouping.
- We have begun to work on Task 3.
  We have demonstrated by Monte Carlo simulations that the GMLE of $\underline{\beta}$ is consistent.
- We have begun to work on Task 4.
  We have demonstrated by Monte Carlo simulations that the TSE of $\underline{\beta}$ is consistent.

E. REPORTABLE OUTCOMES

- A computer program to calculate the GMLE of the baseline survival function S_0 and that of the Cox regression coefficients $\underline{\beta}$.
- A computer program to calculate the TSE of S_0 and that of $\underline{\beta}$.

Both of these computer programs have been made available for the public via the internet site *math.binghamton.edu.ftp/pub/qyu.*

F. CONCLUSIONS

In the first year of our DOD grant, we have successfully completed our research objectives stated in Tasks 1 and 2. In addition, we have begun research work pertaining to Tasks 3 and 4. We have implemented a computer program to compute both the TSE of the baseline survival function and the TSE of the regression coefficients of the Cox regression model under interval censorship. Using Monte Carlo simulations, we have demonstrated that the TSE of the regression coefficients is consistent.

The results which we have established will be useful to breast cancer researchers pursuing chemoprevention intervention trials involving surrogate endpoints biomarkers, and genetic epidemiologists conducting studies on familial aggregation of breast cancer and related cancers.

G. REFERENCES

[1] Cox, D.R. (1972). Regression models and life tables. *J. Roy. Statist. Soc. B,* 34 187-220.

[2] Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, 42 845-854.

[3] Wong, G. Y. C., Bradlow, H. L., Sepkovic, D., Mehl, S., Mailman, J. and Osborne, M. P. (1997). A dose-ranging study of indole-3-carbinol for breast cancer prevention. *Journal of Cellular Biochemistry Supplements* 28/29, 111-116.

[4] Peto, R. (1973). Experimental survival curve for interval-censored data. *Applied Statistics.* 22 86-91.

[5] Turnbull, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser.* B. 38. 290-295.

[6] Groeneboom, P. and Wellner, J.A. (1992). Information bounds and nonparametric maximum likelihood estimation. *Birkhäuser Verlag, Basel.*

[7] Yu, Q. Q., Schick, A., Li, L. X. and Wong, G. Y. C. (1998). Asymptotic properties of the GMLE of a survival function with case 2 interval-censored data. *Statist. & Prob. Let.* 37 223-228.

[8] Yu, Q. Q., Schick, A., Li, L. X. and Wong, G. Y. C. (1998). Asymptotic properties of the GMLE in the case 1 interval-censorship model with discrete inspection times. *Canadian Journal of Statistics* 26 619-627.